CSC2541: Introduction to Causality Lecture 2 - Causal Models

Instructor: Rahul G. Krishnan TA & slides: Vahid Balazadeh-Meresht

September 19, 2022

Outline

Recap - Lecture 1

Review - Bayesian networks Bayesian factorization D-separation Markov properties

Observation & intervention

Causal Bayesian networks Independent mechanisms Formal definition An example Flow of causation and association

Structural causal models

Data generating processes Observational and interventional distributions

Recap - Lecture 2

Lecture 1 Recap

- ▶ What is Causal Inference: It is the study of statistical methods to identify the effect of interventions.
- **Fundamental Problem Of Causal Inference:** We never observe both **potential outcomes** $(Y_1(u), Y_0(u))$ simultaneously.

Estimands of interest:

- 1. Individual Treatment Effect (ITE): What is the effect of an intervention on this individual: $ITE(u) := Y_1(u) Y_0(u)$.
- 2. Average Treatment Effect (ATE): What is the effect of an intervention on a population: ATE := $\mathbb{E}_{u \sim P(u)} [Y_1(u) Y_0(u)]$.
- 3. Conditional Average Treatement Effect: What is the effect of an intervention on a group summarized by covariates that can be conditioned on: $\mathbb{E}[Y_1|X] \mathbb{E}[Y_0|X]$.

Lecture 1 Recap

Problem: The fundamental problem of causal inference makes it challenging to find these estimands without access to an oracle.

Strategy:

- 1. Write down the estimate of interest,
- 2. Make assumptions about the behavior of random variables in the problem,
- 3. Assumptions enable us to write down causal effects using quantities we can estimate from data.

We'll see this strategy arise time and again in this class.

Lecture 1 Recap

Assumptions we covered:

- 1. SUTVA: $Y_{0,1}(u_1) \perp Y_{0,1}(u_k) \forall k \neq 1$
- 2. Consistency: Factual matches the observed outcome
- 3. Ignorability/Exchangeability: Potential outcomes are independent given treatment
- 4. Conditional Ignorability/Exchangeability: Potential outcomes are independent given treatment conditional on covariates [adjustment set]
- 5. Positivity/Overlap: The non-parameteric estimator for ATE requires us to have a positive probability of being assigned treatment or control for each configuration of patient

Positivity Unconfoundedness tradeoff: Including more variables means we're likely to have a valid adjustment set. Comes at the cost of satisfying overlap due to high-dimensionality

Bayesian factorization D-separation Markov properties

Modeling the joint distribution

 $P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

chain rule:
$$P(x_1, x_2, ..., x_n) = \prod_i P(x_i | x_{i-1}, ..., x_1)$$

Bayesian factorization D-separation Markov properties

Modeling the joint distribution

 $P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

chain rule:
$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$$

 pa_i : a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i|x_{i-1}, \ldots, x_1) = P(x_i|pa_i)$

Bayesian factorization D-separation Markov properties

Modeling the joint distribution

 $P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

chain rule:
$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$$

 pa_i : a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i|x_{i-1}, \ldots, x_1) = P(x_i|pa_i)$

$$\begin{split} P(x_1, x_2, x_3, x_4) &= P(x_1) P(x_2 | x_1) P(x_3 | x_2, x_1) P(x_4 | x_3, x_2, x_1) \\ & \text{(Bayesian network factorization} = \text{compact representations)} \end{split}$$

Bayesian factorization D-separation Markov properties

Modeling the joint distribution

 $P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

chain rule:
$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$$

 pa_i : a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i|x_{i-1}, \ldots, x_1) = P(x_i|pa_i)$

$$\begin{split} P(x_1, x_2, x_3, x_4) &= P(x_1) P(x_2 | x_1) P(x_3 | x_1) P(x_4 | x_3, x_2) \\ & (\text{Bayesian network factorization} = \text{compact representations}) \\ & \text{Needs } 2^0 + 2^1 + 2^1 + 2^2 = 9 \text{ parameters} < 2^4 - 1 = 15 \end{split}$$

Bayesian factorization D-separation Markov properties

Graphical models of probabilities



- Directed Acyclic Graph (DAG) ${\mathcal G}$ (Bayesian network)
- P is Markov compatible with \mathcal{G} if our joint distribution admits a factorization compatible with the graph.
- ${\mathcal G}$ describes the conditional independence (CI) structure of distribution P

Bayesian factorization D-separation Markov properties

Conditional Independencies (CI)

What are they?

- Describe structure among the random variables: e.g. what edges do not exist.
- ▶ Provide insight into how information flows within the graph.

Why should we care about CI?

- ▶ We can use this to reduce the storage complexity of joint distribution.
- ▶ Identifying what we should *adjust for* to extract causal effects.

Bayesian factorization D-separation Markov properties

Conditional independence in DAGs

Question

What are the conditional independencies among random variables in a given graph $\mathcal{G}?$

Bayesian factorization D-separation Markov properties

Conditional independence in DAGs

Question

What are the conditional independencies among random variables in a given graph \mathcal{G} ?

We first consider the building blocks of DAGs



Bayesian factorization D-separation Markov properties

Conditional independence - Chain and v-structure

$$X_1 \longrightarrow X_2 \longrightarrow X_3$$

$$P(x_1, x_3 | x_2)$$

$$= \frac{P(x_1, x_2, x_3)}{P(x_2)}$$

$$= \frac{P(x_1)P(x_2|x_1)P(x_3|x_2)}{P(x_2)}$$

= $P(x_1|x_2)P(x_3|x_2)$ (Bayes rule)

$X_1 \perp \!\!\!\perp X_3 | X_2$

Bayesian factorization **D-separation** Markov properties

Conditional independence - Chain and v-structure





 $X_1 \perp\!\!\!\perp X_3 | X_2 \qquad \qquad X_1 \perp\!\!\!\perp X_3$

Bayesian factorization D-separation Markov properties

Conditional independence - Analyzing paths in a graph



Bayesian factorization D-separation Markov properties

Conditional independence - Unblocked paths



Bayesian factorization D-separation Markov properties

Conditional independence - Blocked paths



Bayesian factorization D-separation Markov properties

D-separation

Blocked path Given a DAG \mathcal{G} , a (undirected) path between nodes X and Y is blocked by a set Z iff There is a **chain** $U \rightarrow W \rightarrow V$ or a **fork** $U \leftarrow W \rightarrow V$ on the

▶ There is a **chain** $U \to W \to V$ or a **fork** $U \leftarrow W \to V$ on the path, where $W \in Z$, or

Bayesian factorization D-separation Markov properties

D-separation

Blocked path Given a DAG G, a (undirected) path between nodes X and Y is blocked by a set Z iff There is a chain U → W → V or a fork U ← W → V on the path, where W ∈ Z, or There is a collider U → W ← V on the path, where W ∉ Z and Desc(W) ∉ Z

Bayesian factorization D-separation Markov properties

D-separation

Blocked path Given a DAG G, a (undirected) path between nodes X and Y is blocked by a set Z iff There is a chain U → W → V or a fork U ← W → V on the path, where W ∈ Z, or There is a collider U → W ← V on the path, where W ∉ Z and Desc(W) ∉ Z

D-separation $(X \perp \!\!\!\perp_{\mathcal{G}} Y | Z)$

Given a DAG \mathcal{G} , two sets of nodes X and Y are d-separated by a set Z iff all the paths between nodes of X and Y are blocked by Z

Bayesian factorization D-separation Markov properties

Global and local Markov properties

Idea

Given \mathcal{G} , we can use the Bayes Ball algorithm (Shachter, 1998) to find the conditional independencies in a graph.

Bayesian factorization D-separation Markov properties

Global and local Markov properties

Idea

Given \mathcal{G} , we can use the Bayes Ball algorithm (Shachter, 1998) to find the conditional independencies in a graph.

Global Markov property

Bayesian factorization D-separation Markov properties

Global and local Markov properties

Idea

Given \mathcal{G} , we can use the Bayes Ball algorithm (Shachter, 1998) to find the conditional independencies in a graph.

Global Markov property

Local Markov property

A distribution P satisfies the *local Markov property* w.r.t. a DAG \mathcal{G} if each variable is independent of its nondescendants (in \mathcal{G}) conditioned on its parents.

Bayesian factorization D-separation Markov properties

Global and local Markov properties

Idea

Given \mathcal{G} , we can use the Bayes Ball algorithm (Shachter, 1998) to find the conditional independencies in a graph.

Theorem - Equivalence of Markov properties

Given a distribution P and a DAG \mathcal{G} , if P has a density function, then the followings are equivalent

- 1. P is Markov compatible w.r.t. \mathcal{G}
- 2. P satisfies the global Markov property w.r.t. \mathcal{G}
- 3. P satisfies the local Markov property w.r.t. ${\mathcal G}$

In Markov Random Fields, these properties are shown by the Hammersley-Clifford Theorem.

Bayesian factorization D-separation Markov properties

Observational equivalence

Question

Markov properties relate graphical separation to conditional independencies. Is it possible to have multiple graphs with the same CI structure?

Bayesian factorization D-separation Markov properties

Observational equivalence

Question

Markov properties relate graphical separation to conditional independencies. Is it possible to have multiple graphs with the same CI structure?

Markov equivalence of graphs

Let $\mathcal{M}(\mathcal{G}) := \{P; P \text{ is Markov compatible with } \mathcal{G}\}$. Then, \mathcal{G}_1 and \mathcal{G}_2 are called Markov equivalent if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$.

Bayesian factorization D-separation Markov properties

Observational equivalence

Question

Markov properties relate graphical separation to conditional independencies. Is it possible to have multiple graphs with the same CI structure?

Markov equivalence of graphs

Let $\mathcal{M}(\mathcal{G}) := \{P; P \text{ is Markov compatible with } \mathcal{G}\}$. Then, \mathcal{G}_1 and \mathcal{G}_2 are called Markov equivalent if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$.

Theorem - Observational equivalence

Two DAGs \mathcal{G}_1 and \mathcal{G}_2 are Markov equivalent iff they have the same skeleton and sets of v-structures

Bayesian factorization D-separation Markov properties

Observational equivalence

Theorem - Observational equivalence

Two DAGs \mathcal{G}_1 and \mathcal{G}_2 are Markov equivalent iff they have the same skeleton and sets of v-structures

Bayesian factorization D-separation Markov properties

Observational equivalence

Theorem - Observational equivalence

Two DAGs \mathcal{G}_1 and \mathcal{G}_2 are Markov equivalent iff they have the same skeleton and sets of v-structures



 $P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2,x_3)$

Bayesian factorization D-separation Markov properties

Observational equivalence

Theorem - Observational equivalence

Two DAGs \mathcal{G}_1 and \mathcal{G}_2 are Markov equivalent iff they have the same skeleton and sets of v-structures





 $P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2,x_3) = P(x_2)P(x_1|x_2)P(x_3|x_1)P(x_4|x_2,x_3)$

All these DAGs are observationally valid - They capture the same CI structure

Bayesian factorization D-separation Markov properties

Observational equivalence

Theorem - Observational equivalence

Two DAGs \mathcal{G}_1 and \mathcal{G}_2 are Markov equivalent iff they have the same skeleton and sets of v-structures





 $P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2,x_3) = P(x_3)P(x_1|x_3)P(x_2|x_1)P(x_4|x_2,x_3)$

All these DAGs are observationally valid - They capture the same CI structure

Questions?

Bayesian factorization D-separation Markov properties

Question

Any questions on Bayesian Networks?

Observation and intervention in the kidney stone data - The *do*-operator

Stone size S, treatment T, and recovery rate R. Each item shows P(R = 1|S = s, T = t), i.e., # recovered / #total with T = t, S = s

	Small stone $(S = s)$	Large stone $(S = l)$
T = a Open surgery	81/87	192/263
T = b Percutaneous nephrolithotomy	234/270	55/80

Observation and intervention in the kidney stone data - The do-operator

Stone size S, treatment T, and recovery rate R. Each item shows P(R = 1|S = s, T = t), i.e., # recovered / #total with T = t, S = s

	Small stone $(S = s)$	Large stone $(S = l)$
T = a Open surgery	81/87	192/263
T = b Percutaneous nephrolithotomy	234/270	55/80

Question

Observing a patient with open surgery treatment, what can we infer about their stone size?

Observation and intervention in the kidney stone data - The do-operator

Stone size S, treatment T, and recovery rate R. Each item shows P(R = 1|S = s, T = t), i.e., # recovered / #total with T = t, S = s

	Small stone $(S = s)$	Large stone $(S = l)$
T = a Open surgery	81/87	192/263
T = b Percutaneous nephrolithotomy	234/270	55/80

Question

Observing a patient with open surgery treatment, what can we infer about their stone size?

$$P(S = \text{small}|T = a) = \frac{P(S = \text{small}, T = a)}{P(T = a)} = \frac{87/700}{(87 + 263)/700} \approx 0.25$$
Observation and intervention in the kidney stone data - The do-operator

Stone size S, treatment T, and recovery rate R. Each item shows P(R = 1 | S = s, T = t), i.e., # recovered / #total with T = t, S = s

	Small stone $(S = s)$	Large stone $(S = l)$
T = a Open surgery	81/87	192/263
T = b Percutaneous nephrolithotomy	234/270	55/80

Question

Now, assume we *intervene* on all patients with open surgery treatment. What can we infer about their stone size?

Observation and intervention in the kidney stone data - The do-operator

Stone size S, treatment T, and recovery rate R. Each item shows P(R = 1 | S = s, T = t), i.e., # recovered / #total with T = t, S = s

	Small stone $(S = s)$	Large stone $(S = l)$
T = a Open surgery	81/87	192/263
T = b Percutaneous nephrolithotomy	234/270	55/80

Question

Now, assume we *intervene* on all patients with open surgery treatment. What can we infer about their stone size?

Intuitively, we expect that changes to treatment assignment has no effect on the stone size

$$P(S = \text{small}|\underbrace{do(T = a)}_{\text{intervention}}) = P(S = \text{small}) = \frac{87 + 270}{700} = 0.51$$

- ▶ P(S|T = a): We see (observe) T = a and infer the stone size
- ▶ P(S|do(T = a)): We do (intervene) T = a and infer the stone size
- ► Generally, $P(Y|do(X = x)) \neq P(Y|X = x)$. In the kidney stone data: $P(S = \text{small}|do(T = a)) = P(S = \text{small}) \neq P(S = \text{small}|T = a)$

- ▶ P(S|T = a): We see (observe) T = a and infer the stone size
- ▶ P(S|do(T = a)): We do (intervene) T = a and infer the stone size
- Generally, $P(Y|do(X = x)) \neq P(Y|X = x)$. In the kidney stone data: $P(S = \text{small}|do(T = a)) = P(S = \text{small}) \neq P(S = \text{small}|T = a)$ What about P(R = 1|do(T = a)) and P(R = 1|do(T = b))?

	S = s	S = l
T = a	81/87	192/263
T = b	234/270	55/80

- ▶ P(S|T = a): We see (observe) T = a and infer the stone size
- ▶ P(S|do(T = a)): We do (intervene) T = a and infer the stone size
- Generally, $P(Y|do(X = x)) \neq P(Y|X = x)$. In the kidney stone data: $P(S = \text{small}|do(T = a)) = P(S = \text{small}) \neq P(S = \text{small}|T = a)$ What about P(R = 1|do(T = a)) and P(R = 1|do(T = b))?

P(R = 1 do(T = a))		S = s	S = l
$= P(R_a = 1)$ (Potential outcome)	T = a	81/87	192/263
$=\mathbb{E}[R_a]$	T = b	234/270	55/80

- ▶ P(S|T = a): We see (observe) T = a and infer the stone size
- ▶ P(S|do(T = a)): We do (intervene) T = a and infer the stone size
- Generally, $P(Y|do(X = x)) \neq P(Y|X = x)$. In the kidney stone data: $P(S = \text{small}|do(T = a)) = P(S = \text{small}) \neq P(S = \text{small}|T = a)$ What about P(R = 1|do(T = a)) and P(R = 1|do(T = b))?

$$\begin{split} P(R = 1 | do(T = a)) &= P(R_a = 1) \quad (\text{Potential outcome}) \\ &= \mathbb{E}[R_a] &= \mathbb{E}_S[\mathbb{E}[R | S, T = a]] \quad (\text{G-formula}) \\ &= P(S = s)P(R = 1 | S = s, T = a) + P(S = l)P(R = 1 | S = l, T = a) \\ &= \frac{87 + 270}{700} \cdot \frac{81}{87} + \frac{263 + 80}{700} \cdot \frac{192}{263} \approx 0.832 \end{split}$$

- ▶ P(S|T = a): We see (observe) T = a and infer the stone size
- ▶ P(S|do(T = a)): We do (intervene) T = a and infer the stone size
- Generally, $P(Y|do(X = x)) \neq P(Y|X = x)$. In the kidney stone data: $P(S = \text{small}|do(T = a)) = P(S = \text{small}) \neq P(S = \text{small}|T = a)$ What about P(R = 1|do(T = a)) and P(R = 1|do(T = b))?

P(R = 1 do(T = b))		S = s	S = l
$= P(R_b = 1)$	T = a	81/87	192/263
$=\mathbb{E}[R_b]$ (Potential outcome)	T = b	234/270	55/80
$= \mathbb{E}_{S}[\mathbb{E}[R S, T = \mathbf{b}]] (G-formula)$			
= P(S = s)P(R = 1 S = s, T = b) + P(S = l)P(R = 1 S = l, T = b)			
$=\frac{87+270}{700}\cdot\frac{234}{270}+\frac{263+80}{700}\cdot\frac{55}{80}\approx0.779$			

Can we use potential outcomes to compute interventional distributions?

• Using the potential outcome framework (G-formula), we saw that treatment a is, on average, a better choice

P(R = 1 | do(T = a)) > P(R = 1 | do(T = b))

Can we use potential outcomes to compute interventional distributions?

• Using the potential outcome framework (G-formula), we saw that treatment a is, on average, a better choice

$$P(R = 1 | do(T = a)) > P(R = 1 | do(T = b))$$

Can we use the same G-formula for the following data?

	Normal BP	High/low BP
T = a	81/87	192/263
T = b	234/270	55/80

Can we use potential outcomes to compute interventional distributions?

• Using the potential outcome framework (G-formula), we saw that treatment a is, on average, a better choice

$$P(R = 1 | do(T = a)) > P(R = 1 | do(T = b))$$

Can we use the same G-formula for the following data?

	Normal BP	High/low BP
T = a	81/87	192/263
T = b	234/270	55/80

Since the data is the same, using the same formula will choose treatment a again. But, we saw in lecture 1 that treatment b is better in this case. Why?

Can we use potential outcomes to compute interventional distributions?

• Using the potential outcome framework (G-formula), we saw that treatment a is, on average, a better choice

$$P(R = 1 | do(T = a)) > P(R = 1 | do(T = b))$$

Can we use the same G-formula for the following data?

	Normal BP	High/low BP
T = a	81/87	192/263
T = b	234/270	55/80

- Since the data is the same, using the same formula will choose treatment a again. But, we saw in lecture 1 that treatment b is better in this case. Why?
- Conditional ignorability does not hold BP is not a valid adjustment set $R_a, R_b \not\perp T | BP$ while $R_a, R_b \perp T | S$

Can we use potential outcomes to compute interventional distributions?

• Using the potential outcome framework (G-formula), we saw that treatment a is, on average, a better choice

$$P(R = 1 | do(T = a)) > P(R = 1 | do(T = b))$$

Can we use the same G-formula for the following data?

	Normal BP	High/low BP
T = a	81/87	192/263
T = b	234/270	55/80

- Since the data is the same, using the same formula will choose treatment a again. But, we saw in lecture 1 that treatment b is better in this case. Why?
- Conditional ignorability does not hold BP is not a valid adjustment set $R_a, R_b \not\perp T | BP$ while $R_a, R_b \perp T | S$

▶ It's not always easy to decide what to include in the adjustment set

Can we use potential outcomes to compute interventional distributions?

• Using the potential outcome framework (G-formula), we saw that treatment a is, on average, a better choice

$$P(R = 1 | do(T = a)) > P(R = 1 | do(T = b))$$

Can we use the same G-formula for the following data?

	Normal BP	High/low BP
T = a	81/87	192/263
T = b	234/270	55/80

- Since the data is the same, using the same formula will choose treatment a again. But, we saw in lecture 1 that treatment b is better in this case. Why?
- Conditional ignorability does not hold BP is not a valid adjustment set $R_a, R_b \not\perp T | BP$ while $R_a, R_b \perp T | S$

It's not always easy to decide what to include in the adjustment set
Bayesian networks are a visual tool to better understand adjustment sets to model causal effects

Bayesian networks as data generating processes (DGPs)

Bayesian networks model the conditional independence structure of distribution

Bayesian networks as data generating processes (DGPs)

Bayesian networks model the conditional independence structure of distribution



Bayesian networks as data generating processes (DGPs)

Bayesian networks model the conditional independence structure of distribution



All three graphs are plausible based on data (observationally equivalent)

We are interested in the "real" graph, i.e., the one that describes the real/physical data generating process (Causal order)

Questions?

Question

Any questions on Observations v.s. Interventions?

Independent mechanisms Formal definition An example Flow of causation and association

Bayesian networks as data generating processes (DGP)

Assume that nature generates the data with ordering (T, BP, R)



Independent mechanisms Formal definition An example Flow of causation and association

Bayesian networks as data generating processes (DGP)

Assume that nature generates the data with ordering (T, BP, R)

1. Generate the treatment policy for each patient: P(T)



Independent mechanisms Formal definition An example Flow of causation and association

Bayesian networks as data generating processes (DGP)

Assume that nature generates the data with ordering (T, BP, R)

- 1. Generate the treatment policy for each patient: P(T)
- 2. Generate the blood pressure based on the treatment policy: P(BP|T)



Independent mechanisms Formal definition An example Flow of causation and association

Bayesian networks as data generating processes (DGP)

Assume that nature generates the data with ordering (T, BP, R)

- 1. Generate the treatment policy for each patient: P(T)
- 2. Generate the blood pressure based on the treatment policy: P(BP|T)
- 3. Generate the recovery rate based on the policy and blood pressure: P(R|T, BP)



Independent mechanisms Formal definition An example Flow of causation and association

Bayesian networks as data generating processes (DGP)

Assume that nature generates the data with ordering (T, BP, R)

- 1. Generate the treatment policy for each patient: P(T)
- 2. Generate the blood pressure based on the treatment policy: P(BP|T)
- 3. Generate the recovery rate based on the policy and blood pressure: P(R|T, BP)



For each node X_i in the data generating Bayesian network, $P(x_i|pa_i)$ is called the *mechanism* that generates X_i

How to characterize these "causal" mechanisms?

Independent mechanisms Formal definition An example Flow of causation and association

Characterizing causal mechanisms

Suppose we know the joint distribution P(A, T) of altitude of cities A and their average temperature T. Which one is the cause?

Independent mechanisms Formal definition An example Flow of causation and association

Characterizing causal mechanisms

Case 1: Suppose the causal DGP is $T \to A$ with mechanisms P(T) and $P(A \,|\, T)$

Independent mechanisms Formal definition An example Flow of causation and association

Characterizing causal mechanisms

Case 1: Suppose the causal DGP is $T \to A$ with mechanisms P(T) and $P(A \,|\, T)$



Independent mechanisms Formal definition An example Flow of causation and association

Characterizing causal mechanisms

Case 1: Suppose the causal DGP is $T \to A$ with mechanisms P(T) and P(A|T)



Intervention on T does **not** affect the value of A but **both** mechanisms P(T) and P(A|T) change.

Independent mechanisms Formal definition An example Flow of causation and association

Characterizing causal mechanisms

Case 2: Suppose the causal DGP is $A \to T$ with mechanisms P(A) and P(T|A)

Independent mechanisms Formal definition An example Flow of causation and association

Characterizing causal mechanisms

Case 2: Suppose the causal DGP is $A \to T$ with mechanisms P(A) and P(T|A)



Independent mechanisms Formal definition An example Flow of causation and association

Characterizing causal mechanisms

Case 2: Suppose the causal DGP is $A \to T$ with mechanisms P(A) and P(T|A)



Intervention on A **does** affect the value of T. Also, only **one** mechanism changes (P(A)).

Independent mechanisms Formal definition An example Flow of causation and association

Characterizing causal mechanisms

Case 2: Suppose the causal DGP is $A \to T$ with mechanisms P(A) and P(T|A)



Intervention on A does affect the value of T. Also, only one mechanism changes (P(A)).

A is the cause since intervention in A changes T.

Independent mechanisms Formal definition An example Flow of causation and association

Characterizing causal mechanisms

Case 2: Suppose the causal DGP is $A \to T$ with mechanisms P(A) and P(T|A)



Intervention on A does affect the value of T. Also, only one mechanism changes (P(A)).

A is the cause since intervention in A changes T. Moreover, interventions can only change **one** mechanism in the causal DGP $A \rightarrow T$.

Independent mechanisms Formal definition An example Flow of causation and association

Modularity assumption

Modularity assumption (Independent mechanisms / Autonomy)

A (data generating) Bayesian network has modular mechanisms if intervention on a node X_i only changes the mechanism $P(x_i|pa_i)$

Independent mechanisms Formal definition An example Flow of causation and association

Modularity assumption

Modularity assumption (Independent mechanisms / Autonomy)

A (data generating) Bayesian network has modular mechanisms if intervention on a node X_i only changes the mechanism $P(x_i|pa_i)$

We define a **causal** Bayesian network as a Markov compatible DAG (w.r.t. data distribution) that has modular mechanisms

Independent mechanisms Formal definition An example Flow of causation and association

Modularity assumption

Modularity assumption (Independent mechanisms / Autonomy)

A (data generating) Bayesian network has modular mechanisms if intervention on a node X_i only changes the mechanism $P(x_i|pa_i)$

We define a **causal** Bayesian network as a Markov compatible DAG (w.r.t. data distribution) that has modular mechanisms



Independent mechanisms Formal definition An example Flow of causation and association

Modularity assumption

Modularity assumption (Independent mechanisms / Autonomy)

A (data generating) Bayesian network has modular mechanisms if intervention on a node X_i only changes the mechanism $P(x_i|pa_i)$

We define a **causal** Bayesian network as a Markov compatible DAG (w.r.t. data distribution) that has modular mechanisms



Independent mechanisms Formal definition An example Flow of causation and association

Formal definition of causal Bayesian networks

Causal Bayesian networks

Let P(x) be a probability distribution on a set of variables X and P(x|do(Z = z)) denote the distribution of X after intervention on a subset Z (i.e., setting Z to a constant z).
Independent mechanisms Formal definition An example Flow of causation and association

Formal definition of causal Bayesian networks

Causal Bayesian networks

Let P(x) be a probability distribution on a set of variables X and P(x|do(Z = z)) denote the distribution of X after intervention on a subset Z (i.e., setting Z to a constant z). A DAG \mathcal{G} is **causal Bayesian network compatible with** P iff for every $Z \subseteq X$ and z we have:

Independent mechanisms Formal definition An example Flow of causation and association

Formal definition of causal Bayesian networks

Causal Bayesian networks

Let P(x) be a probability distribution on a set of variables X and P(x|do(Z = z)) denote the distribution of X after intervention on a subset Z (i.e., setting Z to a constant z). A DAG \mathcal{G} is **causal Bayesian network compatible with** P iff for every $Z \subseteq X$ and z we have:

1. P(x|do(Z = z)) is Markov compatible with \mathcal{G}

Independent mechanisms Formal definition An example Flow of causation and association

Formal definition of causal Bayesian networks

Causal Bayesian networks

Let P(x) be a probability distribution on a set of variables X and P(x|do(Z = z)) denote the distribution of X after intervention on a subset Z (i.e., setting Z to a constant z). A DAG \mathcal{G} is **causal Bayesian network compatible with** P iff for every $Z \subseteq X$ and z we have:

- 1. P(x|do(Z = z)) is Markov compatible with \mathcal{G}
- 2. $P(x_i|do(Z = z)) = 1$ for every $X_i \in Z$

Independent mechanisms Formal definition An example Flow of causation and association

Formal definition of causal Bayesian networks

Causal Bayesian networks

Let P(x) be a probability distribution on a set of variables X and P(x|do(Z = z)) denote the distribution of X after intervention on a subset Z (i.e., setting Z to a constant z). A DAG \mathcal{G} is **causal Bayesian network compatible with** P iff for every $Z \subseteq X$ and z we have:

- 1. P(x|do(Z = z)) is Markov compatible with \mathcal{G}
- 2. $P(x_i | do(Z = z)) = 1$ for every $X_i \in Z$
- 3. $P(x_i | pa_i, do(Z = z)) = P(x_i | pa_i)$ for every $X_i \notin Z$

Independent mechanisms Formal definition An example Flow of causation and association

Formal definition of causal Bayesian networks

Causal Bayesian networks

Let P(x) be a probability distribution on a set of variables X and P(x|do(Z = z)) denote the distribution of X after intervention on a subset Z (i.e., setting Z to a constant z). A DAG \mathcal{G} is **causal Bayesian network compatible with** P iff for every $Z \subseteq X$ and z we have:

1. P(x|do(Z = z)) is Markov compatible with \mathcal{G}

Modularity assumption

2.
$$P(x_i|do(Z=z)) = 1$$
 for every $X_i \in Z$

3.
$$P(x_i|pa_i, do(Z = z)) = P(x_i|pa_i)$$
 for every $X_i \notin Z$

Independent mechanisms Formal definition An example Flow of causation and association

How to calculate interventional distributions? - Truncated factorization

Bayesian network factorization

$$P(x_1, x_2, \cdots, x_n) = \prod_i P(x_i | pa_i)$$

Independent mechanisms Formal definition An example Flow of causation and association

How to calculate interventional distributions? - Truncated factorization

Truncated factorization

 $P(x_1, x_2, \cdots, x_n | do(Z = z))$

Independent mechanisms Formal definition An example Flow of causation and association

How to calculate interventional distributions? - Truncated factorization

Truncated factorization

$$P(x_1, x_2, \cdots, x_n | do(Z = z))$$

= $\prod_i P(x_i | pa_i, do(Z = z))$

Markov compatibility (property 1)

Independent mechanisms Formal definition An example Flow of causation and association

How to calculate interventional distributions? - Truncated factorization

Truncated factorization

$$\begin{aligned} P(x_1, x_2, \cdots, x_n | do(Z = z)) \\ &= \prod_i P(x_i | pa_i, do(Z = z)) \end{aligned} \text{Markov compatibility (property 1)} \\ &= \prod_{x_i \in Z} P(x_i | pa_i, do(Z = z)) \prod_{x_i \notin Z} P(x_i | pa_i, do(Z = z)) \end{aligned}$$

Independent mechanisms Formal definition An example Flow of causation and association

How to calculate interventional distributions? - Truncated factorization

Truncated factorization

$$\begin{split} &P(x_1, x_2, \cdots, x_n | do(Z = z)) \\ &= \prod_i P(x_i | pa_i, do(Z = z)) \\ &= \prod_i P(x_i | pa_i, do(Z = z)) \\ &= \prod_{x_i \in Z} P(x_i | pa_i, do(Z = z)) \prod_{x_i \notin Z} P(x_i | pa_i, do(Z = z)) \\ &= 1 \cdot \prod_{x_i \notin Z} P(x_i | pa_i, do(Z = z)) \\ &\qquad \text{Modularity (property 2)} \end{split}$$

Independent mechanisms Formal definition An example Flow of causation and association

How to calculate interventional distributions? - Truncated factorization

Truncated factorization

$$\begin{split} &P(x_1, x_2, \cdots, x_n | do(Z = z)) \\ &= \prod_i P(x_i | pa_i, do(Z = z)) & \text{Markov compatibility (property 1)} \\ &= \prod_{x_i \in Z} P(x_i | pa_i, do(Z = z)) \prod_{x_i \notin Z} P(x_i | pa_i, do(Z = z)) \\ &= 1 \cdot \prod_{x_i \notin Z} P(x_i | pa_i, do(Z = z)) & \text{Modularity (property 2)} \\ &= \prod_{x_i \notin Z} P(x_i | pa_i) & \text{Modularity (property 3)} \end{split}$$

All of these assumes x consistent with z, otherwise it will be zero

E.g.,
$$P(X_1 = 1, X_2 = 2 | do(X_1 = 0)) = 0$$

Independent mechanisms Formal definition $\mathbf{An} example$ Flow of causation and association

Which treatment should we choose?

	Normal BP	High/low BP
T = a	81/87	192/263
T = b	234/270	55/80

Independent mechanisms Formal definition **An example** Flow of causation and association

Which treatment should we choose?

Assuming the causal graph as

	Normal BP	High/low BP
T = a	81/87	192/263
T = b	234/270	55/80



What is P(R = 1 | do(T = a)) and P(R = 1 | do(T = b))?

Independent mechanisms Formal definition **An example** Flow of causation and association

Which treatment should we choose?

Assuming the causal graph as

	Normal BP	High/low BP	
T = a	81/87	192/263	
T = b	234/270	55/80	$(T) \longrightarrow (R)$

What is P(R = 1 | do(T = a)) and P(R = 1 | do(T = b))?

 $P(R = 1 | do(T = \mathbf{a})) = \sum_{t,x} P(R = 1, T = t, BP = x | do(T = \mathbf{a})) \quad \text{marginalization}$

Independent mechanisms Formal definition **An example** Flow of causation and association

Which treatment should we choose?

Assuming the causal graph as

	Normal BP	High/low BP	
T = a	81/87	192/263	
T = b	234/270	55/80	$(T) \longrightarrow (R)$

What is P(R = 1 | do(T = a)) and P(R = 1 | do(T = b))?

 $P(R = 1 | do(T = a)) = \sum_{t,x} P(R = 1, T = t, BP = x | do(T = a))$ marginalization $= \sum_{x} P(R = 1, T = a, BP = x | do(T = a))$ T = bis inconsistent

Independent mechanisms Formal definition **An example** Flow of causation and association

Which treatment should we choose?

Assuming the causal graph as

	Normal BP	High/low BP	
T = a	81/87	192/263	
T = b	234/270	55/80	$T \longrightarrow R$

What is P(R = 1 | do(T = a)) and P(R = 1 | do(T = b))?

 $P(R = 1|do(T = a)) = \sum_{t,x} P(R = 1, T = t, BP = x|do(T = a))$ marginalization $= \sum_{x} P(R = 1, T = a, BP = x|do(T = a))$ T = bis inconsistent $= \sum P(BP = x|T = a)P(R = 1|BP = x, T = a)$

truncated factorization

Independent mechanisms Formal definition **An example** Flow of causation and association

Which treatment should we choose?

Assuming the causal graph as

	Normal BP	High/low BP	
T = a	81/87	192/263	
T = b	234/270	55/80	$(T) \longrightarrow (R)$

What is P(R = 1 | do(T = a)) and P(R = 1 | do(T = b))?

 $P(R = 1 | do(T = a)) = \sum_{t,x} P(R = 1, T = t, BP = x | do(T = a))$ marginalization $= \sum_{x} P(R = 1, T = a, BP = x | do(T = a))$ T = bis inconsistent

$$= \sum_{x} P(BP = x | T = a) P(R = 1 | BP = x, T = a)$$

truncated factorization

$$= P(R = 1 | T = \mathbf{a}) = \frac{81 + 192}{87 + 263} = 0.78$$

Independent mechanisms Formal definition **An example** Flow of causation and association

Which treatment should we choose?

Assuming the causal graph as

	Normal BP	High/low BP	
T = a	81/87	192/263	
T = b	234/270	55/80	$(T) \longrightarrow (R)$

What is P(R = 1 | do(T = a)) and P(R = 1 | do(T = b))?

 $P(R = 1 | do(T = b)) = \sum_{t,x} P(R = 1, T = t, BP = x | do(T = b))$ marginalization $= \sum_{x} P(R = 1, T = b, BP = x | do(T = b))$ T = a is inconsistent

$$= \sum_{x} P(BP = x | T = b) P(R = 1 | BP = x, T = b)$$

truncated factorization

$$= P(R = 1 | T = b) = \frac{234 + 55}{270 + 80} \approx 0.826$$

Independent mechanisms Formal definition $\mathbf{An} example$ Flow of causation and association

Which treatment should we choose?

Assuming the causal graph as

	Normal BP	High/low BP	
T = a	81/87	192/263	
T = b	234/270	55/80	$(T) \rightarrow (R)$

What is P(R = 1 | do(T = a)) and P(R = 1 | do(T = b))?

$$P(R = 1|do(T = b)) = \sum_{t,x} P(R = 1, T = t, BP = x|do(T = b))$$
marginalization
$$= \sum_{x} P(R = 1, T = b, BP = x|do(T = b))$$

T = a is inconsistent

$$= \sum_{x} P(BP = x | T = b) P(R = 1 | BP = x, T = b)$$

truncated factorization

$$= P(R = 1 | T = b) = \frac{234 + 55}{270 + 80} \approx 0.826$$

Here, treatment b is better (and association *is* causation)

27/43

Independent mechanisms Formal definition **An example** Flow of causation and association

Unobserved variables can change everything!

Independent mechanisms Formal definition $\mathbf{An} example$ Flow of causation and association

Unobserved variables can change everything!



Independent mechanisms Formal definition $\mathbf{An} example$ Flow of causation and association

Unobserved variables can change everything!



Independent mechanisms Formal definition \mathbf{An} example Flow of causation and association

Unobserved variables can change everything!



Independent mechanisms Formal definition An exampleFlow of causation and association

Unobserved variables can change everything!

$$P(R = 1|do(T = a))$$

$$= \sum_{x,y} P(S = x)P(BP = y|T = a)P(R = 1|T = a, S = x, BP = y)$$
(truncated factorization - Try this as HW)
$$T$$

Independent mechanisms Formal definition An exampleFlow of causation and association

Unobserved variables can change everything!

What if, in the previous dataset, the stone size had also influenced both T and R but we didn't observe it?

$$P(R = 1|do(T = a))$$

$$= \sum_{x,y} P(S = x)P(BP = y|T = a)P(R = 1|T = a, S = x, BP = y)$$
(truncated factorization - Try this as HW)
$$\neq \sum_{y} P(BP = y|T = a)P(R = 1|T = a, BP = y)$$

$$P(R = 1|do(T = a)) \text{ in the original graph}$$

28/43

Independent mechanisms Formal definition An exampleFlow of causation and association

Unobserved variables can change everything!

What if, in the previous dataset, the stone size had also influenced both T and R but we didn't observe it?

$$P(R = 1|do(T = a))$$

$$= \sum_{x,y} P(S = x)P(BP = y|T = a)P(R = 1|T = a, S = x, BP = y)$$
(truncated factorization - Try this as HW)
$$\neq \sum_{y} P(BP = y|T = a)P(R = 1|T = a, BP = y)$$

$$P(R = 1|do(T = a)) \text{ in the original graph}$$

What to do in the presence of unobserved variable? \rightarrow Lecture 3

Independent mechanisms Formal definition An example Flow of causation and association

When does a node has a causal effect on another one?

Causal effect

A variable (set) Z has causal effect on a (disjoint) variable (set) X if at least for two z,z^\prime

 $P(X|do(Z=z)) \neq P(X|do(Z=z'))$

Independent mechanisms Formal definition An example Flow of causation and association

When does a node has a causal effect on another one?

Causal effect

A variable (set) Z has causal effect on a (disjoint) variable (set) X if at least for two z, z'

 $P(X|do(Z=z)) \neq P(X|do(Z=z'))$

 \blacktriangleright A node Z (in the compatible causal graph) has no causal effect on its non-descendents

Independent mechanisms Formal definition An example Flow of causation and association

When does a node has a causal effect on another one?

Causal effect

A variable (set) Z has causal effect on a (disjoint) variable (set) X if at least for two z, z' $P(X|do(Z = z)) \neq P(X|do(Z = z'))$

 \blacktriangleright A node Z (in the compatible causal graph) has no causal effect on its non-descendents

Proof by induction: $(x_i \text{ is a root node})$ $P(x_i|do(Z = z)) = P(x_i|\emptyset, do(Z = z))$ $= P(x_i)$ (Modularity) $= P(x_i|do(Z = z'))$

Independent mechanisms Formal definition An example Flow of causation and association

When does a node has a causal effect on another one?

Causal effect

A variable (set) Z has causal effect on a (disjoint) variable (set) X if at least for two z, z' $P(X|do(Z = z)) \neq P(X|do(Z = z'))$

 \blacktriangleright A node Z (in the compatible causal graph) has no causal effect on its non-descendents

Proof by induction: $(x_i \text{ is a child node})$ $P(x_i|do(Z = z)) = \sum_{pa_i} P(x_i, pa_i|do(Z = z))$

Independent mechanisms Formal definition An example Flow of causation and association

When does a node has a causal effect on another one?

Causal effect

A variable (set) Z has causal effect on a (disjoint) variable (set) X if at least for two z, z' $P(X|do(Z = z)) \neq P(X|do(Z = z'))$

 \blacktriangleright A node Z (in the compatible causal graph) has no causal effect on its non-descendents

Proof by induction: (x_i is a child node) $P(x_i|do(Z = z)) = \sum_{pa_i} P(x_i, pa_i|do(Z = z))$ $= \sum_{pa_i} P(x_i|pa_i, do(Z = z)) P(pa_i|do(Z = z))$

Independent mechanisms Formal definition An example Flow of causation and association

When does a node has a causal effect on another one?

Causal effect

A variable (set) Z has causal effect on a (disjoint) variable (set) X if at least for two z, z' $P(X|do(Z = z)) \neq P(X|do(Z = z'))$

 \blacktriangleright A node Z (in the compatible causal graph) has no causal effect on its non-descendents

Proof by induction:
$$(x_i \text{ is a child node})$$

 $P(x_i|do(Z = z)) = \sum_{pa_i} P(x_i, pa_i|do(Z = z))$
 $= \sum_{pa_i} P(x_i|pa_i, do(Z = z))P(pa_i|do(Z = z))$
 $= \sum_{pa_i} \underbrace{P(x_i|pa_i, do(Z = z'))}_{\text{Modularity}} \underbrace{P(pa_i|do(Z = z'))}_{\text{Induction step}}$

Independent mechanisms Formal definition An example Flow of causation and association

When does a node has a causal effect on another one?

Causal effect

A variable (set) Z has causal effect on a (disjoint) variable (set) X if at least for two z, z' $P(X|do(Z = z)) \neq P(X|do(Z = z'))$

 \blacktriangleright A node Z (in the compatible causal graph) has no causal effect on its non-descendents

Proof by induction:
$$(x_i \text{ is a child node})$$

 $P(x_i|do(Z = z)) = \sum_{pa_i} P(x_i, pa_i|do(Z = z))$
 $= \sum_{pa_i} P(x_i|pa_i, do(Z = z))P(pa_i|do(Z = z))$
 $= \sum_{pa_i} \underbrace{P(x_i|pa_i, do(Z = z'))}_{\text{Modularity}} \underbrace{P(pa_i|do(Z = z'))}_{\text{Induction step}}$
 $= P(x_i|do(Z = z'))$

Independent mechanisms Formal definition An example Flow of causation and association

Flow of causation v.s. statistical association

A node can only influence its descendents in a causal graph \mathcal{G} ,

- 1. If X is a child of Z in \mathcal{G} : direct cause
- 2. If X is a descendent (and not a child) of Z: *indirect* cause

Independent mechanisms Formal definition An example Flow of causation and association

Flow of causation v.s. statistical association

A node can only influence its descendents in a causal graph \mathcal{G} ,

- 1. If X is a child of Z in \mathcal{G} : direct cause
- 2. If X is a descendent (and not a child) of Z: *indirect* cause



Independent mechanisms Formal definition An example Flow of causation and association

Flow of causation v.s. statistical association

A node can only influence its descendents in a causal graph \mathcal{G} ,

- 1. If X is a child of Z in \mathcal{G} : direct cause
- 2. If X is a descendent (and not a child) of Z: *indirect* cause

Remember d-separation

- Unblocked paths between X_2 and X_5 are (potential) dependencies between X_2 and X_5


Independent mechanisms Formal definition An example Flow of causation and association

Flow of causation v.s. statistical association

A node can only influence its descendents in a causal graph \mathcal{G} ,

- 1. If X is a child of Z in \mathcal{G} : direct cause
- 2. If X is a descendent (and not a child) of Z: *indirect* cause



- Unblocked paths between X_2 and X_5 are (potential) dependencies between X_2 and X_5
- Intervention on X_2 only changes the mechanism $P(X_2|Pa_2) = P(X_2|X_1)$

Independent mechanisms Formal definition An example Flow of causation and association

Flow of causation v.s. statistical association

A node can only influence its descendents in a causal graph \mathcal{G} ,

- 1. If X is a child of Z in \mathcal{G} : direct cause
- 2. If X is a descendent (and not a child) of Z: *indirect* cause

- Unblocked paths between X_2 and X_5 are (potential) dependencies between X_2 and X_5
 - Intervention on X_2 only changes the mechanism $P(X_2|Pa_2) = P(X_2|X_1)$
 - Removing incoming edges to intervened node X_2 : mutilated (or interventional) graph $\mathcal{G}_{\overline{X_2}}$



Independent mechanisms Formal definition An example Flow of causation and association

Flow of causation v.s. statistical association

A node can only influence its descendents in a causal graph \mathcal{G} ,

- 1. If X is a child of Z in \mathcal{G} : direct cause
- 2. If X is a descendent (and not a child) of Z: *indirect* cause

- X_1 X_2 X_3 X_4 X_5
- Unblocked paths between X_2 and X_5 are (potential) dependencies between X_2 and X_5
 - Intervention on X_2 only changes the mechanism $P(X_2|Pa_2) = P(X_2|X_1)$
 - Removing incoming edges to intervened node X_2 : mutilated (or interventional) graph $\mathcal{G}_{\overline{X_2}}$
 - Every unblocked path from X_2 to X_5 in $\mathcal{G}_{\overline{X_2}}$ is a *causal* path (directed paths)

Independent mechanisms Formal definition An example Flow of causation and association

Flow of causation v.s. statistical association

A node can only influence its descendents in a causal graph \mathcal{G} ,

- 1. If X is a child of Z in \mathcal{G} : direct cause
- 2. If X is a descendent (and not a child) of Z: *indirect* cause



- Unblocked paths between X_2 and X_5 are (potential) dependencies between X_2 and X_5
- Intervention on X_2 only changes the mechanism $P(X_2|Pa_2) = P(X_2|X_1)$
- Removing incoming edges to intervened node X_2 : mutilated (or interventional) graph $\mathcal{G}_{\overline{X_2}}$
- Every unblocked path from X_2 to X_5 in $\mathcal{G}_{\overline{X_2}}$ is a *causal* path (directed paths)
- Other unblocked paths in the original graph are backdoor paths

Independent mechanisms Formal definition An example Flow of causation and association

Questions?

Question

Any questions on Causal Bayesian Networks?

Data generating processes Observational and interventional distributions

Modeling data generating processes with equations

• We assumed the data is being generated using (independent) mechanisms $P(x_i|pa_i)$.

Data generating processes Observational and interventional distributions

Modeling data generating processes with equations

▶ We assumed the data is being generated using (independent) mechanisms $P(x_i|pa_i)$. E.g., P(T), P(BP|T), P(R|T, BP) in the kidney data

Data generating processes Observational and interventional distributions

Modeling data generating processes with equations

▶ We assumed the data is being generated using (independent) mechanisms $P(x_i|pa_i)$. E.g., P(T), P(BP|T), P(R|T, BP) in the kidney data

$$R \sim P(R|T = t, BP = x)$$

 $R = f_{x,t}(U)$ where $U \sim \text{Unif}[0, 1]$ and $f_{x,t}(u) = P^{-1}(u|T = t, BP = x)$

Data generating processes Observational and interventional distributions

Modeling data generating processes with equations

▶ We assumed the data is being generated using (independent) mechanisms $P(x_i|pa_i)$. E.g., P(T), P(BP|T), P(R|T, BP) in the kidney data

$$R \sim P(R|T = t, BP = x)$$

 $R = f_{x,t}(U)$ where $U \sim \text{Unif}[0, 1]$ and $f_{x,t}(u) = P^{-1}(u|T = t, BP = x)$

- Any causal mechanism $P(x_i|pa_i)$ can be written as a deterministic function f_i of its direct causes pa_i and some exogenous noise U_i
- We call f_i the law (process) that generates X_i

Data generating processes Observational and interventional distributions

Structural causal models - A mathematical framework to define causal effects

Structural causal model (SCM)

A structural causal model is a tuple $\mathcal{M} = (X, U, F, P_U)$ of

- 1. Endogenous set of variables X, (observed variables)
- 2. Exogenous set of noises U, (unobserved noise)
- 3. Set of functions F, (data generating rules/processes)
- 4. Product distribution P_U over variables in U, i.e., $P_U(u_1, \ldots, u_d) = \prod_{i=1}^d P_U(u_i)$ (noises are independent)

such that for any variable $X_i \in X$, we have an assignment

 $X_i := f_i(PA_i, U_i)$ But not $f_i(PA_i, U_i) := X_i$

for some $PA_i \subseteq X \setminus \{X_i\}$, $U_i \in U$, and $f_i \in F$. We call the elements of PA_i direct causes of X_i .

Data generating processes Observational and interventional distributions

SCMs induce causal graphs

For each SCM \mathcal{M} , we can construct a (unique) graph \mathcal{G} by drawing edges from each direct cause in PA_i to X_i

$$\begin{aligned} X_1 &:= f_1(U_1) \\ X_2 &:= f_2(X_1, U_2) \\ X_3 &:= f_3(X_2, U_3) \\ X_4 &:= f_4(X_1, X_3, U_4) \\ U_1, \dots, U_4 \text{ are jointly independent} \end{aligned}$$



causal graph

• We assume \mathcal{G} is acyclic (no feedback loop in assignments)

$$X_1 := f_1(X_2, U_1)$$
$$X_2 := f_2(X_1, U_2)$$

Generating "observational" distribution with SCMs

How to generate data from an SCM?

- 1. Consider a topological order of endogenous variables X_1, \ldots, X_d (since the assignments are acyclic)
- 2. Sample from exogenous noises $u_1, \ldots, u_d \sim P_U$
- 3. Generate samples x_1, \ldots, x_d by assignments $x_i = f_i(pa_i, u_i)$

Each X_i can be written as a unique function of noises $(U_k)_{k \in An_i}$ that belong to ancestors of X_i , i.e.,

$$X_i = g_i((U_k)_{k \in An_i})$$

Generating "observational" distribution with SCMs

How to generate data from an SCM?

- 1. Consider a topological order of endogenous variables X_1, \ldots, X_d (since the assignments are acyclic)
- 2. Sample from exogenous noises $u_1, \ldots, u_d \sim P_U$
- 3. Generate samples x_1, \ldots, x_d by assignments $x_i = f_i(pa_i, u_i)$

Each X_i can be written as a unique function of noises $(U_k)_{k \in An_i}$ that belong to ancestors of X_i , i.e.,

$$X_i = g_i((U_k)_{k \in An_i})$$

Observational distribution

An SCM \mathcal{M} induces a unique distribution over endogenous variables X_1, \ldots, X_d , which we call the observational distribution of \mathcal{M} and denote it by $P_X^{\mathcal{M}}$, or simply P.

Data generating processes Observational and interventional distributions

Generating "interventional" distribution with SCMs

Remember the independent mechanisms assumption: intervention on a variable X_i can only change the mechanism $P(x_i|pa_i)$

We can use SCMs to formally define interventions

Interventional distribution

Consider an SCM \mathcal{M} . An intervention on a variable X_i (or multiple variables) is replacing the assignment $X_i := f_i(PA_i, U_i)$ with a new assignment

$$X_i := \hat{f}(\overline{PA}_i, \, \hat{U}_i)$$

We call the induced distribution of the new SCM an interventional distribution and denote it by $P_X^{\mathcal{M}}\left(\cdot | do\left(X_i := \hat{f}(\overline{PA}_i, \hat{U}_i)\right)\right)$. If $\hat{f}(\overline{PA}_i, \hat{U}_i)$ is a constant value c, we simply write it as $P(\cdot| do(X_i = c))$.

Data generating processes Observational and interventional distributions

Types of intervention - soft intervention

Soft interventions: $\overline{PA}_i = PA_i$, i.e., only the mechanism changes but direct causes remain active

before intervention

$$X_1 := \mathcal{N}(0, 1)$$

$$X_2 := \mathcal{N}(0, 1)$$

$$X_3 := X_1 + X_2 + \mathcal{N}(0, 1)$$



Data generating processes Observational and interventional distributions

Types of intervention - soft intervention

Soft interventions: $\overline{PA}_i = PA_i$, i.e., only the mechanism changes but direct causes remain active

after intervention on X_3

$$\begin{split} X_1 &:= \mathcal{N}(0, 1) \\ X_2 &:= \mathcal{N}(0, 1) \\ X_3 &:= \frac{X_1^2 + X_2^2}{2} + \texttt{Unif}(0, 1) \end{split}$$



Data generating processes Observational and interventional distributions

Types of intervention - hard intervention

▶ Hard intervention: $\overline{PA}_i \neq PA_i$

before intervention

$$X_1 := \mathcal{N}(0, 1)$$

$$X_2 := \mathcal{N}(0, 1)$$

$$X_3 := X_1 + X_2 + \mathcal{N}(0, 1)$$



Data generating processes Observational and interventional distributions

Types of intervention - hard intervention

▶ Hard intervention: $\overline{PA}_i \neq PA_i$

after intervention on X_3

$$X_1 := \mathcal{N}(0, 1)$$
$$X_2 := \mathcal{N}(0, 1)$$
$$X_3 := \mathbf{2X}_1 + \mathcal{N}(0, 1)$$



Data generating processes Observational and interventional distributions

Types of intervention - hard intervention

▶ Hard intervention: $\overline{PA}_i \neq PA_i$

after intervention on X_3

$$X_1 := \mathcal{N}(0, 1)$$
$$X_2 := \mathcal{N}(0, 1)$$
$$X_3 := c$$



▶ Atomic intervention is a type of hard intervention, where $X_i := c$ for some constant value $c \to$ mutilated graph $\mathcal{G}_{\overline{X_2}}$

Data generating processes Observational and interventional distributions

Types of intervention - hard intervention

▶ Hard intervention: $\overline{PA}_i \neq PA_i$

after intervention on X_3

$$X_1 := \mathcal{N}(0, 1)$$
$$X_2 := \mathcal{N}(0, 1)$$
$$X_3 := c$$



- Atomic intervention is a type of hard intervention, where $X_i := c$ for some constant value $c \to$ mutilated graph $\mathcal{G}_{\overline{X_2}}$
- ▶ We previously defined causal Bayesian networks only using atomic interventions (see slide 25). SCMs give us more flexibility in defining interventions
- ▶ The causal graph corresponding to an SCM \mathcal{M} is a causal Bayesian network compatible with $P^{\mathcal{M}}$

Data generating processes Observational and interventional distributions

Example - good predictors are not always causes

Consider the following SCM:

 $X_1 := U_{X_1}$ $Y := X_1 + U_Y$ $X_2 := Y + U_{X_2}$

 $U_{X_1}, U_Y \sim \mathcal{N}(0, 1)$ $U_{X_2} \sim \mathcal{N}(0, \mathbf{0.1})$

Data generating processes Observational and interventional distributions

Example - good predictors are not always causes

Consider the following SCM:

$$\begin{split} X_1 &:= U_{X_1} \\ Y &:= X_1 + U_Y \\ X_2 &:= Y + U_{X_2} \end{split}$$

 $U_{X_1}, U_Y \sim \mathcal{N}(0, 1)$ $U_{X_2} \sim \mathcal{N}(0, \mathbf{0.1})$



Observational distribution for X_2 and Y

Data generating processes Observational and interventional distributions

Example - good predictors are not always causes

Consider the following SCM:

$$\begin{split} X_1 &:= U_{X_1} \\ Y &:= X_1 + U_Y \\ X_2 &:= Y + U_{X_2} \end{split}$$

$$U_{X_1}, U_Y \sim \mathcal{N}(0, 1)$$
$$U_{X_2} \sim \mathcal{N}(0, \mathbf{0.1})$$

We train two linear models to predict Y:

1. $\hat{Y}_1 = \theta_1 X_1$: $\mathbb{E} \left[\| \hat{Y}_1 - Y \|_2^2 \right] \approx 1$ 2. $\hat{Y}_2 = \theta_2 X_2$: $\mathbb{E} \left[\| \hat{Y}_2 - Y \|_2^2 \right] \approx 0.1$



Observational distribution for X_2 and Y

Data generating processes Observational and interventional distributions

Example - good predictors are not always causes

Now, we intervene on X_2 :

 $X_1 := U_{X_1}$ $Y := X_1 + U_Y$ $X_2 := U_{X_2}$

 $U_{X_1}, U_Y \sim \mathcal{N}(0, 1)$ $U_{X_2} \sim \mathcal{N}(0, 1)$

Data generating processes Observational and interventional distributions

Example - good predictors are not always causes

Now, we intervene on X_2 :

 $X_1 := U_{X_1}$ $Y := X_1 + U_Y$ $X_2 := U_{X_2}$

 $U_{X_1}, U_Y \sim \mathcal{N}(0, 1)$ $U_{X_2} \sim \mathcal{N}(0, 1)$

 X_2 is not a good predictor for Y anymore (independent of Y)



Interventional distribution for X_2 and Y

Data generating processes Observational and interventional distributions

Questions?

Question

Any questions on Structural Causal Models?

Recap

- Bayesian networks Compact representations of joint probability distributions.
- Conditional Independencies Blocked and Unblocked paths characterize the flow of association.
- D-separation and (global/local) Markov properties Characterize conditional independence in a graph.
- Observational equivalence We cannot distinguish graphs that have the same skeleton and same v-structures from data.

Recap

- do-operator Operator that corresponds to an intervention on a random variable.
- ▶ Independent mechanisms (or Modularity) Intervention on a node only changes the mechanism associated with that node.
- Causal Bayesian Networks G is causal BN if the interventional distribution is Markov compatible with it and it satisfies modularity.
- Analyzing (directed) paths in a Causal Bayesian Network lets us assess the flow of causation.
- Structural Causal Models Functional representation of causal process that generates the data (more flexibility than Bayesian network).
- ▶ Good predictors need not be causal!